

Inférence basée sur la vraisemblance empirique. Le cas d'échantillonnage complexes

Yves G. Berger

University of Southampton, UK

La vraisemblance empirique classique a été principalement développée dans le cas d'observations indépendantes et identiquement distribuées (iid). Pour des données issues d'un sondage les observations ne sont pas iid. Dès lors, la vraisemblance empirique classique ne peut pas directement être appliquée. Pour remédier à ce problème, la pseudo-vraisemblance empirique a été développée. Cette approche tient compte du plan de sondage par l'intermédiaire d'un terme correctif: l'effet de sondage. Cette approche possède certains désavantages. Nous proposons une nouvelle approche basée sur la vraisemblance empirique qui tient compte du plan d'échantillonnage. La vraisemblance empirique proposée est différente de la pseudo-vraisemblance empirique et ne dépend pas de l'effet de sondage.

L'approche que nous proposons n'est pas une approche basée sur un modèle de super-population. Il s'agit d'une approche purement basée sur le plan de sondage. Les paramètres considérés sont des paramètres de population finie. L'approche proposée récupère certains avantages de l'approche de vraisemblance classique tout en étant basé sur le plan d'échantillonnage et sans supposer un modèle de super-population.

L'approche proposée permet d'estimer de manière consistante et asymptotiquement optimales des paramètres qui sont solutions d'équations estimantes (par exemple: moyennes, totaux, quantiles, paramètres de régressions,...). L'avantage de l'approche proposée est qu'elle permet de construire des intervalles de confiance sans devoir estimer la variance. Ces intervalles de confiance ne nécessitent pas la normalité de l'estimateur ponctuelle. La linéarisation ou le ré-échantillonnage (jackknife ou bootstrap) n'est pas non plus nécessaire, même dans le cas où le paramètre d'intérêt n'est pas linéaire. Cette approche donne de bons intervalles de confiance même si les données sont fortement asymétriques, ou lorsque la linéarisation ne donne pas de bon estimés de variance, ou lorsqu'on s'intéresse à des domaines. L'approche proposée permet aussi d'estimer des paramètres de modèles de régression généralisée (par exemple, la régression logistique) et de tester s'ils sont significatifs (sous une approche basée sur le plan d'échantillonnage). L'information auxiliaire peut être tenue en compte de manière naturelle et très simple.

Nous montrerons des résultats de simulations pour des moyennes, des quantiles des paramètres de régression (simple et logistique). Le bootstrap est une autre méthode qui peut être utilisée pour dériver des intervalles de confiance. L'approche proposée est plus simple à mettre en œuvre que le bootstrap. Comme le bootstrap, l'approche proposée ne repose pas sur des dérivations analytiques. Nos simulations montrent que le bootstrap et la linéarisation peuvent donner de mauvais taux de recouvrement. L'approche proposée donne de meilleurs taux de recouvrement.

L'approche proposée a été appliquées sur le sondage de pauvreté de l'Union Européenne (European Union Survey on Income and Living Conditions, EU-SILC). Nous comparerons les intervalles de confiance proposés avec les intervalles de confiance basés sur la linéarisation et le bootstrap.